Anomaly Analysis and Diagnosis for Co-located Datacenter Workloads in the Alibaba Cluster

*Presented by Rui Ren*

Institute of Computing Technology, CAS

2019-11-14
Motivation

- Co-located workloads in Datacenter
  - Curse of resource utilization and quality of service
    - Resource utilization
    - Response time
  - Alibaba tried to deploy batch jobs and latency-critical online services on the same machines.
Motivation

- Alibaba Cluster Trace: contains online services and batch jobs.
- The data is provided to address the challenges Alibaba face in idcs where online services and batch jobs are co-allocated:
  - 1. Workload characterizations.
  - 2. New algorithms to assign workload.
  - 3. Online service and batch jobs scheduler cooperation.
Goals

- Recent studies:
  - Analyzing the characteristics from the perspective of imbalance phenomenon, co-located workloads (how the co-located workloads interact and impact each other), the elasticity and plasticity of semi-containerized cloud.

- However, discovering the cluster anomalies quickly is important, which helps to locate bottlenecks, troubleshoot problems and improve utilization.

We perform a deep analysis on the released Alibaba co-located trace dataset, from the perspective of anomaly analysis and diagnosis, and try to reveal several insights!
Trace Overview

1) Resource Data:
   b) Container Resource Usage: container_event.csv, container_usage.csv.

2) Workload Data:
   batch task.csv and batch instance.csv

Raw Data Preprocessing

Supplement the missing data and filter the abnormal data.

- For the missing machine 149, 602 and 930 in file server usage.csv, all resource data is completed with 0.
- There are several missing resource usage records on 335 machines, and there missing data are filled up by linear interpolation method.
Raw Data Preprocessing

Aggregate all the container-level, batch-level and server-level resource usage statistics by the machine id and recording interval (300s).

- Generating container-level resource usage data.
- Generating batch-level resource usage data.
- Generating server-level resource usage data.

\[
\begin{align*}
RT(bi)_{m,I_x} &= t(bi)_{end} - t(bi)_{start} \quad (t(bi)_{start} \geq t_x \& t(bi)_{end} \leq t_{x+1}) \\
RT(bi)_{m,I_x} &= t(bi)_{end} - t_x \quad (t(bi)_{start} \leq t_x \& t(bi)_{end} \leq t_{x+1}) \\
RT(bi)_{m,I_x} &= t_{x+1} - t(bi)_{start} \quad (t(bi)_{start} \geq t_x \& t(bi)_{end} \geq t_{x+1}) \\
RT(bi)_{m,I_x} &= t_{x+1} - t_x \quad (t(bi)_{start} \leq t_x \& t(bi)_{end} \geq t_{x+1})
\end{align*}
\]
Distributions of Resource Utilization

The box-and-whisker plot showing CPU usage and memory usage distributions:

- The aggregated CPU usage of online containers is lower than that of batch tasks.
- The aggregated memory usage of online containers is higher than that of batch tasks.

It implies that most batch jobs are computational tasks, and the online container services (long-running jobs) are more memory-demanding.
Distributions of Resource Utilization

The resource usage heatmap of online containers.

- There are no running online containers from the range of machine 132 to 151, machine 418 to 553.
- During the tracing interval, the resource utilization (CPU usage and memory usage) of online containers is relatively stable.
Distributions of Resource Utilization

The resource usage heatmap of batch tasks.

There are no running batch tasks from 14.7h in several machine regions: the region of machine 95 to 127, 275 to 296, 753 to 760, 830 to 906.

The resource utilization is not as stable as that of long-running jobs, especially the memory usage is fluctuating.

The online containers are the long-running jobs with more memory-demanding, so the memory usage is relatively stable; while the memory usage of batch jobs is fluctuating, for most batch tasks are short jobs.
Anomaly Analysis

Abnormal node discovery: Isolation Forest (iForest)

If one machine’s anomaly score is smaller, the probability that it is an abnormal node is higher.

81 machines have anomaly scores that are less than 0.

Fig. 5. The anomaly score.
Abnormal cause analysis

Unbalanced Co-located Workload Distribution

- Based on the number of batch tasks and online containers on machines, all machines can be classified into 8 workload distribution categories.
Abnormal cause analysis

- 8 workload distribution categories:

<table>
<thead>
<tr>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
<th>Type 5</th>
<th>Type 6</th>
<th>Type 7</th>
<th>Type 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>956</td>
<td>9</td>
<td>170</td>
<td>11</td>
<td>2</td>
<td>155</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

- Average cosine similarity of all nodes for each workload distribution category:

<table>
<thead>
<tr>
<th>Type</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
<th>Type 5</th>
<th>Type 6</th>
<th>Type 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity</td>
<td>99.17%</td>
<td>99.19%</td>
<td>98.05%</td>
<td>98.23%</td>
<td>99.64%</td>
<td>98.98%</td>
<td>99.17%</td>
</tr>
</tbody>
</table>

- The co-located workload distribution is unbalance: the resource utilization is different between different categories;
- The resource utilization in the same workload distribution category is very similar.
Abnormal cause analysis

- Skew of co-located workload resource utilization
  - Resource utilization ratio:
    \[
    Cpu\_ratio_{m, \lambda} = \frac{Cpu(bi)_{m, \lambda}}{Cpu(ci)_{m, \lambda}}
    \]
    \[
    Mem\_ratio_{m, \lambda} = \frac{Mem(bi)_{m, \lambda}}{Mem(ci)_{m, \lambda}}
    \]
  - The larger the ratio is, the higher the resource utilization of batch jobs is.
  - The lower the ratio is, the higher the resource utilization of the online containers is.
Abnormal cause analysis

- The histogram and cumulative distribution function (CDF) curve of different ratio ranges

- 74.4% of Cpu ratio is greater than 1, which means the batch tasks are CPU-intensive workloads with higher cpu utilization.
Abnormal cause analysis

- The histogram and cumulative distribution function (CDF) curve of different ratio ranges

- 76.59% of Mem ratio is less than 1, which means the memory occupied by the batch tasks is not high, and the online containers have higher memory requirements and utilization.
Abnormal cause analysis

- System Failures
  - The timeline of soft errors on different machines
Abnormal cause analysis

- Failed Instances
  - The failed instance number of machines:

![Graph showing failed instances](image)
Abnormal cause analysis

- Failed Instances
  - Top 10 machines that have the most failed instances:

<table>
<thead>
<tr>
<th>$m$</th>
<th>679</th>
<th>680</th>
<th>673</th>
<th>341</th>
<th>823</th>
<th>664</th>
<th>232</th>
<th>1006</th>
<th>536</th>
<th>1040</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Num(FIns)$</td>
<td>823</td>
<td>471</td>
<td>444</td>
<td>433</td>
<td>392</td>
<td>347</td>
<td>307</td>
<td>306</td>
<td>299</td>
<td>297</td>
</tr>
</tbody>
</table>

- The batch instance failed on a node is common, and the Fuxi JobMaster can process these failures based on its fault tolerance mechanism;
- If there are a lot of failed batch instances on a node, which means some states of this node may be not suitable for batch tasks.
### Abnormal Cases Study

#### Top 25 abnormal nodes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Node</th>
<th>Anomaly Score</th>
<th>Category</th>
<th>Unbalanced Workload Distribution</th>
<th>Skew Co-workload Utilization</th>
<th>Softerrors</th>
<th>Failed Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>930</td>
<td>-0.1709</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>602</td>
<td>-0.1709</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1075</td>
<td>-0.1527</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>550</td>
<td>-0.1525</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>372</td>
<td>-0.1524</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>478</td>
<td>-0.1522</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>983</td>
<td>-0.1508</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>924</td>
<td>-0.1506</td>
<td>Type 2</td>
<td>No workloads</td>
<td>Frequent softerrors</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>676</td>
<td>-0.1451</td>
<td>Type 1</td>
<td>No workload</td>
<td>Heavier online services</td>
<td>289</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>481</td>
<td>-0.1428</td>
<td>Type 2</td>
<td>No workload</td>
<td>Heavier online services</td>
<td>823</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>679</td>
<td>-0.1395</td>
<td>Type 1</td>
<td>No workload</td>
<td>Heavier online services</td>
<td>444</td>
<td>101</td>
</tr>
<tr>
<td>12</td>
<td>851</td>
<td>-0.1223</td>
<td>Type 4</td>
<td>No batch jobs</td>
<td>Softerrors</td>
<td>166</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>673</td>
<td>-0.1198</td>
<td>Type 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>993</td>
<td>-0.1105</td>
<td>Type 6</td>
<td>Unbalanced batch tasks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>618</td>
<td>-0.0928</td>
<td>Type 8</td>
<td>Unbalanced batch tasks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>556</td>
<td>-0.0833</td>
<td>Type 4</td>
<td>No batch jobs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>689</td>
<td>-0.0827</td>
<td>Type 5</td>
<td>Unbalanced workloads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>401</td>
<td>-0.0826</td>
<td>Type 5</td>
<td>Unbalanced workloads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>275</td>
<td>-0.07879</td>
<td>Type 6</td>
<td>Unbalanced batch tasks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>763</td>
<td>-0.0774</td>
<td>Type 4</td>
<td>No batch jobs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>149</td>
<td>-0.0724</td>
<td>Type 3</td>
<td>No online services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1039</td>
<td>-0.0720</td>
<td>Type 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>800</td>
<td>-0.0662</td>
<td>Type 4</td>
<td>No batch jobs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1069</td>
<td>-0.0646</td>
<td>Type 4</td>
<td>No batch jobs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>949</td>
<td>-0.0629</td>
<td>Type 4</td>
<td>No batch jobs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

we conclude the possible anomalies causes of co-located cluster:

- The Unbalanced co-located workload distribution has a great impact on the resource utilization of cluster nodes, which leads to abnormal nodes.
- Skew co-located workload resource utilization also results in several abnormal nodes.
- Frequent system failures have a large impact on system status.
Anomaly Analysis and Diagnosis for Co-located Datacenter Workloads in the Alibaba Cluster

Q & A?

Thank You!